Modeling the semantic coordination of speech and gesture under cognitive and linguistic constraints

Kirsten Bergmann, Sebastian Kahl, and Stefan Kopp

SFB 673, CITEC, Faculty of Technology, Bielefeld University, Germany
{kbergman,skahl,skopp}@techfak.uni-bielefeld.de

Abstract. This paper addresses the semantic coordination of speech and gesture, a major prerequisite when endowing virtual agents with convincing multimodal behavior. Previous research has focused on building rule- or data-based models specific for a particular language, culture or individual speaker, but without considering the underlying cognitive processes. We present a flexible cognitive model in which both linguistic as well as cognitive constraints are considered in order to simulate natural semantic coordination across speech and gesture. An implementation of this model is presented and first simulation results, compatible with empirical data from the literature are reported.

Keywords: Speech, gesture, generation, cognitive modeling, semantic coordination

1 Introduction

Intelligent virtual agents are required to be able to express themselves convincingly and autonomously. There is a growing body of evidence demonstrating the importance of nonverbal expressivity for this purpose. Especially co-speech gestures play a major role in providing agents with convincing communication skills, because they are an integral part of human communication, inseparably intertwined with speech [24]. It has, for instance, been demonstrated that a virtual agent's use of gestures might enhance the perceived verbal eloquence [7] as well as users' ratings of the agent, mostly in terms of competence [3]. State of the art agent systems, however, still "remain a long way from matching the complexity and subtlety of real-life nonverbal behavior" [21, p. 80]. One major reason for this is that many aspects of how humans use and produce speech and gestures in combination are not fully understood, yet.

There is considerable theoretical disagreement regarding the process by which semantic coordination between speech and gesture is achieved. This involves two major issues, information distribution and information packaging. Information distribution means that the two modalities, although expressing the same underlying idea, do not convey necessarily identical aspects of it: Gestures can be found to be *redundant* with the information encoded verbally (e.g., 'round cake'

+ gesture depicting a round shape), to *supplement* it (e.g., 'cake' + round gesture), or even to *complement* it (e.g., 'looks like this' + round gesture). Closely related to how information is distributed across modalities, is information packaging: How much information is put into a multimodal utterance? When are messages rather split into several parts? With regard to our example this means that the above utterance might also be split into two clauses, e.g., 'there is a cake' followed by 'and it is round', whereby both clauses might be accompanied by a gesture, e.g. a supplementary pointing gesture with the first clause, and a redundant shape-depicting gesture accompanying the second clause.

Empirical evidence suggests that both cognitive and linguistic constraints are involved in the process of meaning coordination. However, a concrete and comprehensive picture of how variations in meaning coordination arise under those constraints is still missing. In recent work we developed a cognitive model for the production of speech and gestures [19]. This model already covers *formulation* processes determining the surface form of speech and gestures as well as the *conceptualization* process by which meaning is structured, portioned and distributed across the two modalities. The latter is based on activation-spreading within dynamically shaped multimodal memories. We have shown how this model can simulate and explain different cases of information distribution under cognitive constraints.

In this paper we present an extension of the model to also consider linguistic constraints and, consequently, information packaging. We start with a review of empirical findings, followed by a discussion of related work. We then present our model which provides a detailed cognitive account of how meaning is dynamically organized and coordinated across speech and gesture. Finally, we present new modeling results demonstrating that the model can account for empirical findings as reported in the literature.

2 Background and Related Work

2.1 Information distribution across speech and gesture

A couple of studies have investigated how the frequency and nature of gesturing, including its coordination with speech is influenced by *cognitive factors*. Bavelas et al. [2], for instance, found that speakers are more likely to produce non-redundant gestures when their addressees could see them, as opposed to when their gestures are not visible and hence less essential for their partners. Bergmann and Kopp [4] report results from an analysis of natural co-verbal gesturing in direction-giving, indicating that supplementary gestures are more likely in cases of problems of speech production (e.g. disfluencies) or when the information conveyed is introduced into the dialogue (and thus conceptualized for the first time). In line with this, recent work has suggested that speakers indeed produce more gestures at moments of relatively high load on the conceptualization process for speaking [17], in particular on the linearization and the focusing components of conceptualization [25].

Hostetter and Alibali [13] report findings suggesting that speakers who have stronger visual-spatial skills than verbal skills produce higher rates of gestures than other speakers. In a later study, Hostetter and Alibali [14] found that the speakers with high spatial skills also produced a higher proportion of non-redundant gestures than other speakers, whereas verbal-dominant speakers tended to produce such gestures more in case of speech disfluencies. The authors hypothesize that "*non-redundant gesture-speech combinations occur because mental images are more active in speaker's minds at the moment of speaking than are verbal codes*" [p.45]. Taken together this suggests that non-redundant gesture-speech combinations are the result of speakers having both strong spatial knowledge and weak verbal knowledge simultaneously, and avoiding the effort of transforming the one into the other.

2.2 Information packaging for speech and gesture

Empirical evidence investigating speech-gesture information packaging suggests that gestures are influenced by *linguistic constraints* in terms of conceptual, syntactic, and lexical structure of concomitant speech. In a cross-linguistic study Kita and Özyürek [15] demonstrated that the packaging of content for gestures parallels linguistic information packaging. Speakers of Japanese, Turkish and English had to re-tell cartoon events for which their languages provide differing means of encoding. English speakers, for example, used the verb 'swing' for a character's action, encoding an arc-shaped trajectory, while Turkish and Japanese speakers employed a trajectory-neutral, change-of-location predicate such as 'move'. Gestures followed this packaging in a way that Japanese and Turkish speakers were more likely to produce straight gestures, whereas most English speakers produced arced gestures. In another cartoon-event, the character rolled down a hill. Again, speakers of English typically described this by combining manner and path of the movement in a single clause (e.g. "he rolled down"), accompanied by a single gesture encoding both semantic features. In contrast, Turkish and Japanese speakers encode manner and path separately in two clauses (e.g. "he descended as he rolled") and also used two separate gestures for these two features.

Evidence along the same line comes from a study on language acquisition [27]: Advanced L2 speakers of English typically encoded manner and path information in one clause and their gestures followed, whereas speakers at lower proficiency levels typically used two-clause constructions in speech, accompanied by separate gestures for manner and path. A subsequent study [16] showed that this effect also occurs when L1 speakers of English are forced to produce one- or two-clause descriptions of manner and path.

Kita and Özyürek [15] proposed an explanatory models for these empirical findings. This account explicitly incorporates the idea that language shapes iconic gestures such that the content of a gesture is determined by three factors: (1) the speaker's communicative intention, (2) action schemata selected on the basis of features of imagined or real space, and (3) bidirectional interactions between speech and gesture production processes. The latter takes place at the

level of conceptualization, i.e. the organization of meaning. An additional interaction between the speech formulator and the (preverbal) message generator is assumed to allow for feedback from grammatical or phonological encoding to the conceptualizer and thus to gesture.

2.3 Computational models of speech and gesture production

Computational modeling of speech-gestural communicative behavior for virtual agents mostly focused on how gesture use is constrained by linguistic features, while cognitive constraints remained completely disregarded. Existing approaches basically fall into two groups depending on how they bring speech and gestures together: *rule-based* and *data-driven* approaches. Among the rule-based models, the BEAT system [10], for instance, was based on behavior generators in which generation rules extracted from empirical data were implemented. This way the framework considered linguistic information such as information structure for the selection of predefined gesture specifications. A similar approach was taken in the Nonverbal Behavior Generator (NVBG) [22]. The system analyzes the syntactic and semantic structure of surface texts and takes the affective state of the virtual agent into account to generate appropriate nonverbal behaviors. Based on a study from the literature and a video analysis of emotional dialogues, the authors developed a list of nonverbal behavior generation rules. In the REA architecture [9] gestures were lexicalized like words and selected using a lexical choice algorithm and incorporated directly into natural language generation. In particular, it implemented rules of information distribution to account for the fact that speech and gestures are sometimes redundant and sometimes complementary. The NUMACK account [20] followed the same strategy by using an integrated microplanner (SPUD) to compose multimodal utterances. Extended with a flexible gesture planner (instead of using a static set of predefined gestures), gestures were dynamically incorporated into SPUD's resources and utilized in grammatically pre-determined ways.

Data-driven models have been adopted mainly to account for individual or cultural style of co-verbal gesturing. Neff et al. [26] developed a system generating gesture animations for novel text by using speaker-specific gesture profiles that were created from a corpus of communicative behavior. Based on these profiles, the system made probabilistic generation choices conditioned upon the previously performed gesture and the input text tagged with theme, rheme, and focus. Similarly, Endrass et al. [12] proposed a corpus-driven method of generating gestures in a culture-specific way that accompany virtual agent's verbal utterances. The frequency of gestures and gesture-types, the correlation of gesture-types and speech-acts as well as the expressivity of gestures have been analyzed in the two cultures of Germany and Japan and integrated into a generation model.

Apart from models developed for virtual agents, [8] have proposed a cognitive modeling attempt for the production of speech and gestures, using the cognitive architecture ACT-R [1]. This account draws on two major assumptions: (1) on

the claim that language representations include some irreducibly spatial components; (2) on the idea that language processing is based on constructions which consist of both semantic and syntactic components. The authors assume these constructions to prescribe spatial representations for what they call *linguistic spatial gestures* and which they assume to provide only "*little information not included in the accompanying language*" [p.14].

In sum, the fact that the virtual agent community has undertaken multiple different efforts to develop production accounts in which gestures are constrained by their linguistic context clearly shows the relevance of such models. This is particularly emphasized by the recent trend of cross-culturally (or cross-linguistically) employed agents. However, existing generation systems basically solve this by combining or switching between static models, rule-based or data-based, explicitly developed for a specific language, culture or individual. What we present in this paper is a flexible cognitive model that simulates natural semantic coordination of speech and iconic gesture, accounting for information distribution and packaging under dynamically arising or changing linguistic and cognitive constraints.

3 A Cognitive Model

In order to investigate to what extent information distribution and information packaging across modalities can be explained by cognitive and linguistic constraints we developed a model based on activation-based processing on multimodal memory [19]. This account is embedded in a larger production model that comprises three stages: conceptualization, where a *message generator* and an *image generator* work together to select and organize information to be encoded in speech and gesture, respectively; formulation, where a *speech formulator* and a *gesture formulator* determine appropriate verbal and gestural forms for this; *motor control* and *articulation* to finally execute the behaviors. Motor control, articulation, and formulation have been subject of earlier work [5].

3.1 Multimodal Memory

The central component in our model is a multimodal memory which is accessible by modules of all processing stages. We assume that language production requires a preverbal message to be formulated in a symbolic-propositional representation that is linguistically shaped [23] (SPR, henceforth). During conceptualization the SPR, e.g. a function-argument structure denoting a spatial property of an object, often needs to be extracted from visuo-spatial representations (VSR), e.g. the mental image of this object. We assume this process to involve the invocation and instantiation of memorized supramodal concepts (SMC, henceforth), e.g. the concept 'round' which links the corresponding visuo-spatial properties to a corresponding propositional denotation. Fig. 1 illustrates the overall relation of these tripartite multimodal memory structures.

To realize the VSR and part of the SMC, we employ a model of visuo-spatial imagery called *Imagistic Description Trees* (IDT) [28]. The IDT model was designed, based on empirical data, to cover the meaningful visuo-spatial features in shape-depicting iconic gestures. Each node in an IDT contains an imagistic description which holds a schema representing the shape of an object or object part. Important aspects include (1) a tree structure for shape decomposition, with abstracted object schemas as nodes, (2) extents in different dimensions as an approximation of shape, and (3) the possibility of dimensional information to be underspecified. The latter occurs, e.g., when the axes of an object schema cover less than the three dimensions of space or when an exact dimensional extent is left open but only a coarse relation between axes like "dominates" is given. This allows to represent the visuo-spatial properties of SMCs such as 'round', 'left-of' or 'longish'. Applying SMC to VSR is realized through graph unification and similarity matching between object schemas, yielding similarity values that assess how well a certain SMC applies to a particular visuo-spatially represented entity (cf. Fig. 1). SPR are implemented straight forward as predicate-argument sentences.

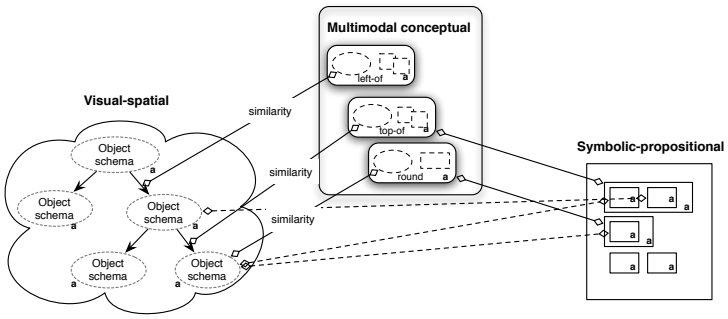


Fig. 1. Tripartite multimodal memory structure involved in speech and gesture production.

3.2 Formulators and generators

The *message generator* has to pre-package the activated SPR information in a way that the *speech formulator* can produce an appropriate sentence. We employ an LTAG-based sentence planner for speech formulation. To make sure that all facts necessary to generate a verbal utterance are available, the *message generator* applies networks that reflect the options of verbalization provided in the *speech formulator*'s LTAG grammar (this conforms the view that the conceptualizer learns to anticipate the formulator's abilities [23]).

The *image generator* retrieves visuo-spatial information about the object to be described, from activated and salient VSR and SMC entries in memory. It is in charge of unifying this information into an imagistic description, from which the *gesture formulator* can derive a gesture form specification (based on Bayesian decision networks learned from empirical data [5]). For instance, information about shape might be combined with information about the object's size or position (as encoded in the IDT representation of VSR entries). Depending on the knowledge encoded here, the *gesture formulator* is able to plan a shape-depicting gesture or rather a localizing gesture.

3.3 Overall production process

Fig. 2 shows an outline of the overall production architecture. We assume that conceptualization consists of cognitive processes that operate upon the multimodal memory structures and are constrained by principles of memory retrieval, which can be modeled by principles of activation spreading [11]. As assumed in cognitive architectures like ACT-R [1], activations float dynamically, spread across linked entities (in particular via SMCs), and decay over time. Activation of more complex SMCs are thereby assumed to decay slower than activations in VSR or SPR. For implementation details see [19].

Formulation-based reinforcement Information distribution is explained *via* a mechanism of formulation-based reinforcement: Propositions encoded in the speech formulators' first formulation suggestions result in reinforced activation of this concept in SPR memory, and thus increased activation of the associated concept in VSR. This is possible due to the fact that the speech formulator links words and semantics in terms of SPR entries. Similarly, imagistic representations encoded by the *gesture formulator* also result in respectively reinforced activations in VSR memory, spreading over to SPR memory. In result, multimodal coordination in terms of information distribution emerges from the local choices the generators and formulators take based on the activation dynamics in multimodally linked memory representations. Note that as activation is dynamic, which features are selected becomes dependent on the time of retrieval and, thus, available cognitive resources. Redundant speech and gesture, then, result from focused activation of supramodally linked mental representations, whereas non-redundant speech and gesture arise when activations in VSR and SPR scatter over entries that are not connected via SMC concepts.

Goal-based reinforcement and activation decrease Our account of information packaging rests upon two ideas: (1) goal-based intensification and (2) activation decrease when information is successfully conveyed. Production always starts with the *image generator* and *message generator* inducing activations in modal entries, evoked by a communicative intention (goal) such as "introduce churchwindow-1". Upon retrieval, the *generators* independently select features and pass them to the respective *formulator*. The communicative intention is held

available in the generators as long as it has not been fulfilled, i.e., until all memory entries involved have been put into words or gestures. Notably, this need not take place in only one utterance. It might as well be that the communicative intention is split into two or more clauses. As soon as an utterance is passed to realization, the activation of associated concepts in VSR and SPR is decreased. Only those parts of the communicative intention that have not been realized so far are kept activated. As this might be a single concept, the process of goal-based reinforcement now ensures that enough contextual information is available for the generators to initiate the generation of an entire clause or sentence.

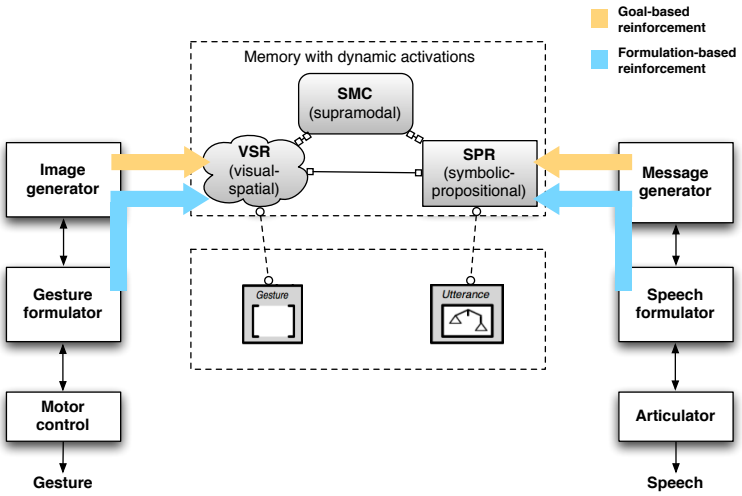


Fig. 2. Overall production architecture.

4 Modeling Results

Our model has been implemented into our speech-gesture production architecture as a fully operational prototype that can directly be connected to standard behavior realization engines driving a virtual character. As a first exploration, we report results from a study on how (1) cognitive resources, specifically processing time, and (2) linguistic constraints in terms of limited verbalization ability affect meaning coordination in the produced multimodal utterances. We first describe the emerging production process in detail and then report quantitative results.

The production process is always initiated by setting the communicative intention "introduce churchwindow-1". Upon receiving this goal, the *image generator* activates visuo-spatial imagery of the church window in VSR, and the *message generator* activates symbolic representations of non-spatial semantic

concepts in SPR. These activations spread through memory and lead to invocation of SMCs for, e.g., 'round' (bound to churchwindow-1) and 'at-top-of' (the church-tower), as well as instantiation of the corresponding SPR entries.

SMCs along with their linked entries in VSR and SPR attain highest and most slowly decaying activation values. Multimodal memory processing is based on so-called update cycles. Per cycle entry activations are being calculated, spread across associations and new associations are created if connections are sufficiently active. The frequency with which a memory cycle is being triggered can be modified with respect to the speed at which the system is required to run. After a preset number of processing cycles, both generators retrieve modality-specific information from memory with a probability depending on current activation values, leading to 'round' and 'at-top-of' concepts being encoded in speech and gesture in a less coordinated/aligned way. If there is enough processing time available, the *message generator* starts to collect concepts for the preverbal message and re-activates those entries being retrieved and used for *speech formulation*: the contents expressed either verbally or gesturally tend to converge. This results in well-coordinated multimodal representations when the modality-specific formulators finally start with their generation work. Thus, it is more likely that both generators receive the same information about shape and position of the entity. Accordingly, the *speech formulator* is enabled to plan a sentence like "The church has a round window at the top" accompanied by a shape depicting gesture like drawing the shape of the window in the air, or a static posturing gesture where the hands becoming a model of the circular shape. As the position of the entity is also available, the gesture would be performed in that part of gesture space.

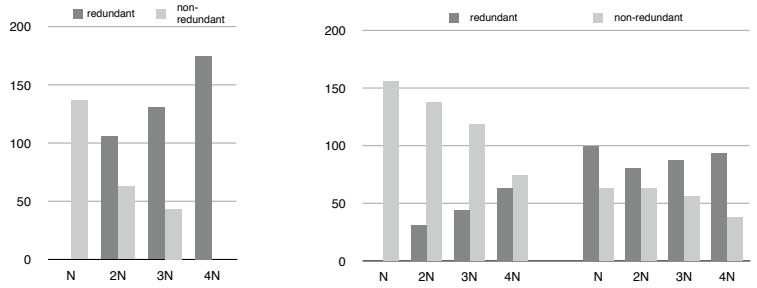
Generation under cognitive constraints If the cognitive processing time available is short, goal-reinforced content coordination is restricted and the multimodal representations are less well coordinated when the modality-specific formulators start with their generation work. So the *message generator* may retrieve only information about the salient shape of the window, but not about its position relative to other entities. Thus, a sentence like "The church has a round window" gets formulated. The *image generator*, on the other hand, may receive information about the entity's position as well. This might result in gestures like the ones described previously–encoding both, shape and position information.

Generation under linguistic constraints To simulate cross-linguistic variation as described in [15], we assume as linguistic constraint that a speaker's grammar does not provide a noun phrase construction of the type 'DET ADJ NN', i.e., we modify the linguistic abilities embodied in the *message generator* and the *speech formulator*. The *speech formulator* is hence not able to return a one-sentence solution for all propositions. Instead, it proposes a solution like "The church has a window (at the top)" which does not convey 'round' bound to churchwindow-1. Depending on whether the accompanying gesture encodes the feature 'round', which is possible but not very likely as there is no formulation-based reinforcement of the SMC 'round', the deactivation process after utterance realization

leaves the 'round' feature at a high activation level so that the generators again initiate an goal-based reinforcement of memory structures. Now the *speech formulator* builds a message plan for an utterance containing the 'round'-feature, resulting in a verbalization like 'the window is round'. Due to formulation-based activation of the shape property 'round' by the *message generator*, a shape-depicting gesture is likely to be planned to accompany this second utterance.

4.1 Quantitative results

To quantify these observations, we ran simulation experiments in which we manipulated the time available (in terms of memory update cycles) and analyzed the resulting gestures for their semantic content and semantic coordination with speech in terms of redundancy/non-redundancy. In the simulations we constrained, first, processing time as a cognitive resource by forcing the system after a particular amount of cycles (N, 2N, 3N and 4N cycles) to realize an utterance based on the current memory state. Second, we manipulated the system's verbalization capabilities. In the 'full grammar' (FG) condition the 'DET ADJ NN' construction was available; in the 'limited grammar' (LG) condition the verbal ability of the *message generator* and *speech formulator* was impaired by making the 'DET ADJ NN' construction unavailable. In this condition the communicative intention had to be realized in two clauses. We ran the model 100 times in each condition.



(a) Full grammar (one clause)

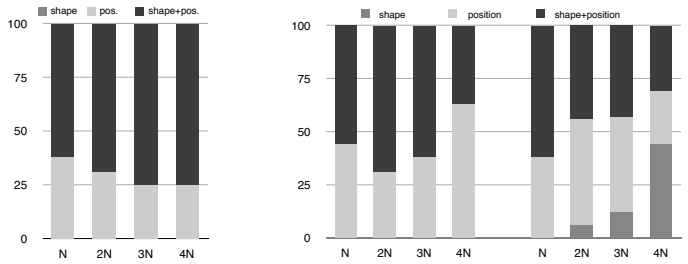
(b) Limited grammar (two consecutive clauses)

Fig. 3. Amount of semantic gesture features encoded redundantly vs. non-redundantly with speech in 100 simulations under manipulation of time available (N, 2N, 3N and 4N memory update cycles).

Figure 3 shows that non-redundant (supplementary) gestures dominate in those runs with stricter cognitive constraints, while redundant ones become more likely when processing time available is increased. As in the LG condition the communicative intention is realized *via* two separate clauses, the amount of

verbally encoded information per clause is lower than in the FG condition. Accordingly, the accompanying gestures are more likely to be non-redundant per se.

Figure 4 shows the effect verbalization ability on semantic coordination. In the FG condition, gestures tend to convey both verbally encoded semantic features, position and shape. The more processing time is available, the more likely it is that gestures express both semantic features (in redundancy with speech). By contrast, in the LG condition, conflated gestures become less likely over time. Rather, gestures accompanying the first clause (*without* verbally encoded shape information) tend to express position information only, while the gestures accompanying the second clause (*with* verbally encoded shape information) tend to express shape information only, the more processing time is available. Moreover, the amount of conflated gestures conveying both semantic features is higher in the FG condition as compared to the LG condition, especially when utterances are produced with more processing time.



(a) Full grammar (one clause)

(b) Limited grammar (two consecutive clauses)

Fig. 4. Semantics encoded by gestures in 100 simulation under manipulation of time available (N, 2N, 3N and 4N memory update cycles).

5 Conclusion

We have presented an extended model to explain semantic coordination of speech and gesture in terms of (1) how visuo-spatial and symbolic-propositional memory structures are dynamically organized, (2) how these multimodal memory structures affect activation-spreading, and (3) how this interacts with modality-specific processes of conceptualization and formulation. The mechanisms of formulation-based and goal-based reinforcement of memory entries in combination with an activation decrease of successfully conveyed information enable to simulate two major empirical findings reported in literature. First, our model allows to simulate cross-linguistic data according to which gestural representations follow linguistic encoding patterns. When verbal capabilities allow for a compact

verbalization of the communicative intention, gestures also tend to conflate the very semantics. When, by contrast, linguistic constraints require semantics to be separated into two clauses, this is reflected in gesture. Second, we have shown in simulation that the model also offers a natural account for the finding that non-redundant gestures are more likely when conceptualization load is high, based on the assumption that memory-based cross-modal coordination consumes resources (memory, time), and is reduced or compromised when such resources are limited.

This exemplifies how a flexible, cognitive account like ours can help to go beyond explicit modeling of the speech-gesture relationship. This way, we can improve human-agent interaction such that it progresses towards intuitive and human-like communication, and we can contribute to understanding the cognitive phenomena by making hypotheses testable in terms of predictions that can be explored in computational simulations as well as in appropriately set up empirical experiments. Future work will therefore be directed to further extend and evaluate the model regarding several respects.

First, we need to extend the model towards evaluation principles to decide upon the moment when utterance planning can be terminated. In other words, when is a multimodal utterance 'good enough' to be realized? This issue is closely related to the question whether a speech-gesture ensemble encodes the communicative intention adequately. Is it, for instance, sufficient to encode shape information only gesturally (as in 'there is a church window' + shaping gesture)? Or do we need a follow-up clause encoding the shape information verbally? A cost-based "good-enough" generation will enable the system to make such decisions autonomously, e.g., contingent upon addressee presence (cf. [29]) or addressee feedback.

Second, we aim to investigate speech-gesture use beyond single ensembles, by considering larger discourse units including sequences of contributions as well as multi-part gestures. Here we need to consider alignment effects: Speakers adapt their utterances with respect to their own or their interlocutor's previous communicative behavior. We have already analyzed such gestural alignment empirically in [6] and found that a speaker's own gestures influence each other more than the gestures the interlocutor performs. Moreover, not all gesture features seem to be equally subject to this effect. For a first sketch of how to account for these findings in a larger perception-action architecture see [18]. Moreover, there is the necessity of planning sequences of several consecutive utterances which also requires to consider the context in terms of what has been said and gesticulated before.

Finally, one might consider cognitive constraints not only as processing time, as we did in our initial modeling attempt here, but also in terms of other factors occurring in human cognition like spatial skills (cf. [13, 14]) or conceptualization load (cf. [25]). A systematical manipulation of such model parameters with analyses of the resulting behavior will provide further insights into the production of multimodal utterances and a more comprehensive picture of how different factors and mechanisms act in concert to produce overt communicative behavior.

6 Acknowledgements

This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 “Alignment in Communication” and the Center of Excellence 277 “Cognitive Interaction Technology” (CITEC).

References

1. Anderson, J., Bothell, D., Byrne, M., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychological Review 111(4), 1036–1060 (2004)
2. Bavelas, J., Kenwood, C., Johnson, T., Philips, B.: An experimental study of when and how speakers use gestures to communicate. Gesture 2(1), 1–17 (2002)
3. Bergmann, K., Eyssel, F., Kopp, S.: A second chance to make a first impression? how appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time. In: Proceedings of IVA 2012. Springer, Berlin/Heidelberg (2012)
4. Bergmann, K., Kopp, S.: Verbal or visual: How information is distributed across speech and gesture in spatial dialog. In: Proceedings of SemDial2006. pp. 90–97 (2006)
5. Bergmann, K., Kopp, S.: GNetIc—Using Bayesian decision networks for iconic gesture generation. In: Proceedings of IVA 2009, pp. 76–89. Springer, Berlin/Heidelberg (2009)
6. Bergmann, K., Kopp, S.: Gestural alignment in natural dialogue. In: Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci 2013). pp. 1326–1331. Cognitive Science Society, Austin, TX (2012)
7. Bergmann, K., Kopp, S., Eyssel, F.: Individualized gesturing outperforms average gesturing–evaluating gesture production in virtual humans. In: Proceedings of IVA 2010. pp. 104–117. Springer, Berlin/Heidelberg (2010)
8. Breslow, L., Harrison, A., Trafton, J.: Linguistic spatial gestures. In: Proceedings of Cognitive Modeling 2010. pp. 13–18 (2010)
9. Cassell, J., Stone, M., Yan, H.: Coordination and context-dependence in the generation of embodied conversation. In: Proceedings of the First International Conference on Natural Language Generation (2000)
10. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: The behavior expression animation toolkit. In: Proceedings of SIGGRAPH 2001. pp. 477–486. New York, NY (2001)
11. Collins, A.M., Loftus, E.F.: A spreading-activation theory of semantic processing. Psychological Review 82(6), 407–428 (1975)
12. Endrass, B., Damian, I., Huber, P., Rehm, M., André, E.: Generating culture-specific gestures for virtual agent dialogs. In: Proceedings of IVA 2010. pp. 329–335. Springer, Berlin/Heidelberg (2010)
13. Hostetter, A., Alibali, M.: Raise your hand if you’re spatial—relations between verbal and spatial skills and gesture production. Gesture 7, 73–95 (2007)
14. Hostetter, A., Alibali, M.: Cognitive skills and gesture-speech redundancy. Gesture 11(1), 40–60 (2011)
15. Kita, S., Özyürek, A.: What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. Journal of Memory and Language 48, 16–32 (2003)

16. Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., Ishizuka, T.: Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. Language and Cognitive Processes 22, 1212–1236 (2007)
17. Kita, S., Davies, T.S.: Competing conceptual representations trigger co-speech representational gestures. Language and Cognitive Processes 24(5), 761–775 (2009)
18. Kopp, S., Bergmann, K.: Automatic and strategic alignment of co-verbal gestures in dialogue. In: Wachsmuth, I., de Ruiter, J., Jaecks, P., Kopp, S. (eds.) Alignment in Communication: Towards a New Theory of Communication, chap. 6. John Benjamins, Amsterdam, NL (in press)
19. Kopp, S., Bergmann, K., Kahl, S.: A spreading-activation model of the semantic coordination of speech and gesture. In: Proceedings of the 35th Annual Conference of the Cognitive Science Society (CogSci 2013). Cognitive Science Society, Austin, TX (in press)
20. Kopp, S., Tepper, P., Ferriman, K., Striegnitz, K., Cassell, J.: Trading spaces: How humans and humanoids use speech and gesture to give directions. In: Nishida, T. (ed.) Conversational Informatics, pp. 133–160. John Wiley, New York (2007)
21. Kraemer, N., Bente, G.: Personalizing e-learning. the social effects of pedagogical agents. Educational Psycholoy Review 22, 71–87 (2010)
22. Lee, J., Marsella, S.: Nonverbal behavior generator for embodied conversational agents. In: Proceedings of IVA 2006. pp. 243–255. Springer, Berlin (2006)
23. Levelt, W.J.M.: Speaking: From intention to articulation. MIT Press (1989)
24. McNeill, D., Duncan, S.: Growth points in thinking-for-speaking. In: Language and gesture, pp. 141–161. Cambridge University Press, Cambridge, UK (2000)
25. Melinger, A., Kita, S.: Conceptualisation load triggers gesture production. Language and Cognitive Processes 22(4), 473–500 (2007)
26. Neff, M., Kipp, M., Albrecht, I., Seidel, H.P.: Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics 27(1), 1–24 (2008)
27. Özyürek, A.: Speech-gesture relationship across languages and in second language learners: Implications for spatial thinking and speaking. In: Proceedings of the 26th Boston University Conference on Language Development. pp. 500–509 (2002)
28. Sowa, T., Kopp, S.: A cognitive model for the representation and processing of shape-related gestures. In: Proc. European Cognitive Science Conference (2003)
29. Swets, B., Jacovina, M.E., Gerrig, R.J.: Effects of conversational pressures on speech planning. Discourse Processes 50(1), 23–51 (2013)